DOCUMENT RESUME

ED 428 114 TM 029 585

AUTHOR Dodds, Jeffrey
TITLE Writing Good Tests for Student Grading or Research Purposes: Some Basic Precepts and Principles.
PUB DATE 1999-01-21
NOTE 23p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Educational Research; *Grading; Higher Education; Item Analysis; *Item Banks; Objective Tests; *Test Construction; Test Items; Test Use; Test Validity

ABSTRACT

Basic precepts for test development are described and explained as they are presented in measurement textbooks commonly used in the fields of education and psychology. The five building blocks discussed as the foundation of well-constructed tests are: (1) specification of purpose; (2) standard conditions; (3) consistency; (4) validity; and (5) practicality. When these foundations are established, the actual test can be constructed taking each of these areas into account. Before the test items are constructed, it is important to consider issues related to item difficulty. True/false, multiple-choice, matching, short-answer, and essay test items can be considered for specific purposes. Once the initial pool of test items is assembled, item revision, pretests, and item analyses can be conducted. (Contains 16 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 428 114

Running Head: WRITING GOOD TESTS

Writing Good Tests for Student Grading or Research Purposes: Some Basic Precepts and Principles

Jeffrey Dodds

Texas A&M University 77843-4225

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Jeffrey Dodds

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029585

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 21, 1999.

BEST COPY AVAILABLE

Abstract

The integrity of our assessments and research endeavors largely turns upon the quality of our tests. This paper elaborates and explains basic precepts for test development as these precepts are presented in the measurement textbooks commonly used within the fields of education and psychology.

Many of us as educators regularly develop or revise tests of cognitive ability or achievement for various grading purposes. Many of us also develop cognitive tests for use in educational research. Of course, the integrity of our assessments or our research endeavors turns upon the quality of our tests. Test quality, in turn can be impacted by our knowledge of and adherence to established principles of test construction.

This paper elaborates and explains some basic precepts and principles for test development as these are presented in the commonly used measurement textbooks within education and psychology. The compilation will derive from books by such authors as Crocker and Algina (1986), Wiersma and Jurs (1990), Gay (1990), Brown (1983), Sax (1989), and Thorndike, Cunningham, Thorndike and Hagen (1991).

The Importance of Good Testing Procedures

Writing good tests can be demanding, but is nevertheless important. As teachers and researchers increasingly come to understand that tests are not reliable (scores are) (Thompson, 1994), and as "reliability generalization" methods (Vacha-Haase, 1998) are increasingly used, the difficult challenges involved in test construction are increasingly being acknowledged. Use of time-proven precepts and principles can improve the prospects for successful test development.

Thorndike et al. (1991) identified three reasons why the test construction procedures used by most teachers are less than optimal. First of all, few teachers receive much training to

construct good tests in that most teacher-preparation programs require only a minimal amount of course work on this very important topic. Second, follow-up studies have shown that teachers do not retain much of what they learned about test construction and are reluctant to use what they have learned. This reluctance may be partially due to the fact that test construction concepts can be difficult to understand and be very time consuming to employ (Thorndike et al., 1991). A third reason teachers may be reluctant to follow proper test construction procedures is that analyses of item properties and score reliability require knowledge of more difficult computations. Although even a basic understanding of the concepts of the mean and median would allow teachers to see how the typical or average student performed on a test, Gullickson and Ellwein (1985) found that of the primary and secondary teachers they surveyed, only 12% could compute the median and 13% could compute the mean.

The Foundation for Good Tests

Brown (1983) identified five specific elements in the foundation of well constructed tests, namely, specification of purpose, standard conditions, consistency, validity, and practicality. These elements can be viewed as the building blocks in our construction process and will allow us as test writers to reach the goals we are trying to achieve.

The first building block in the foundation of our tests is the specification of purpose. This is a concept which we will explore in greater detail later in the paper. It will be shown

that specifying (a) the construct the test is designed to measure, (b) how the results are going to be used and (c) who will take the test all contribute to the direction we take in the test construction process.

The second building block is the establishment of standard conditions. This is a fundamental way to control for error and hone the accuracy of our scores. By standardizing conditions all along the way in the construction process, we can control for error in the test development stage, in the administration of the test, and in the test scoring.

The next block in our foundation-building process is the concept of consistency with regard to test scores. Unless the test we have constructed will produce consistent scores, the scores will not have much value. This leads us to our fourth block in the foundation, validity.

In order for the test scores to be interpretable, they must be valid. By valid, we mean that the scores represent the construct they were designed to measure, and nothing else. If we have produced a well-constructed test with items of proper difficulty, the validity will be enhanced. Validity may also be enhanced by increased heterogeneity of the group being measured. Again, just like with our standard conditions, subtle individual factors can effect validity.

Lastly, issues of practicality and efficiency must be built into the process. This means we have to consider the time, money and qualifications needed to administer, score, and interpret the

test. Ultimately we want to use the simplest procedures possible while maintaining the highest test quality (Brown, 1983).

Types and Classifications of Tests

The types and classifications of tests vary widely from performance tests, which measure maximum or typical performance, to self-report instruments, such as questionaires, surveys, and interviews. There are also instruments that measure intelligence, personality, aptitude or achievement. In each of these cases, it is important to understand how the test will be referenced. This may be done either by norming or by using a criterion.

Norm-referenced tests are used to make inferences about how much a student has learned in comparison with others, so the decisions being made are "relative" decisions. Usually the norm-referenced test is intended to yield only an overall score. These tests are broader in content than the criterion-referenced tests and direct inferences are not made about which objectives have been mastered by given students.

Criterion-referenced tests are used for "absolute" decisions, such as "Has this student learned the specific course content?". The student's performance on each objective must be assessed at a level of reliability that will permit conclusions about whether the student has achieved mastery. This means the items associated with an objective must, theoretically, be samples from all possible items that could measure that objective (Wiersma & Jurs, 1990).

We will look at some additional concepts regarding these two

types of tests later in this paper. For the purposes of further discussion, we will now look at the construction process in terms of paper-and-pencil tests that measure cognitive ability.

Initial Stages in the Construction Process

Now that we have the foundation of the test in place, we can begin the test's actual construction. The first step is to determine the purpose of the test in terms of who will be tested with the measure and what constructs will be measured. This will vary according to the measurement of knowledge and behavior from the cognitive and psychological domains. We must also simultaneously consider what will be gained from the testing information and how the results will be used.

The second step is to identify a plan for the test. This can be accomplished with the use of a test blueprint or a specification table. In order to use specification tables and test blueprints it is necessary to understand some of the philosophy behind their usage, both of which are based on categories of the Taxonomy of Educational Objectives: The Cognitive Domain (Bloom, Englehart, Furst, Hill & Krathwohl, 1956). The two taxonomies most commonly referred to with regard to test construction are the *cognitive* and *affective* domains of behavior.

The cognitive domain consists of six levels. Level 1, Knowledge, involves the test taker's recall, memorization and recognition of previously learned material like dates, people, and terminology. Level 2, Comprehension, focuses on the test taker's understanding, not just memorization. For instance, using an item

that asks a child to circle all the even numbers from a list would require their understanding of what an even number is, not just that 2 or 4 are even numbers. At level 3, Application, test takers are asked to apply their understanding. An example of an Application test item would be "Compute the standard deviation and variance from a group of scores." Level 4, Analysis, deals with the ability to break down a problem into its basic elements and identify relationships that exist between them. An example of an item testing at this level would be, "Differentiate between a classroom achievement test and a standardized achievement test in terms of what each measures and how each is used." Level 5, Synthesis, involves the ability to combine elements into a unique whole, creating something new. An item testing at this level might ask a student to "devise a plan to reduce the federal deficit." Test items at the sixth level, Evaluation, ask test takers to make a judgment based on reasoning, like making judgments on the value of an idea.

Gronlund (1971) reminded test writers that any test is only a sample of the many possible items that could be included to test what a student has learned. All students are expected to know thousands of facts but are tested on only a limited number of them. The same is true of the number of situations they understand or the problem-solving skills they develop. Each area can be tested with a limited number of items. Therefore, in the case of each area of content and each specific learning outcome we, as teachers, are only selecting a sample of student performance and

accepting it as evidence of achievement in that area. This is why it is so important to use a table of specifications or a test blueprint in the test construction process. We want to develop as representative a sample as possible. Utilizing this taxonomy allows us to develop test items that measure higher mental processes. A major flaw in many teacher-made tests is that they test only at the knowledge level.

The specification table usually takes the form of a two-way grid with major content areas listed in one margin and cognitive processes in the other. The table serves several purposes. First, it helps us to determine how many and what sort of items need to be written. Second, at the end of the test construction process we will be able to check to see if the final form of the test matches the table or test plan. In this way we can see if our items adequately sample the domain we want to cover.

Crocker and Algina (1986) stated that by writing test items according to specifications they will be interchangeable. This procedure is related to another one we will look at in more detail later, the assembly of a pool of test items (writing more items than will actually be included in the test). Some educators suggest developing a test blueprint before any actual instruction occurs. This way an instructor has a clear idea of what concepts should be taught and students will have an idea of the relative emphasis placed on contents and skills. Both the table of specifications and the test blueprint delineate objectives measured, item characteristics, and level of mastery. They also

help us to avoid bias and redundancy of items.

Item Difficulty

Before we construct the actual test items, it is important to consider the function of item difficulty in both norm-referenced and criterion-referenced tests. The difficulty of test items on criterion-referenced tests is determined by the specific learning task to be measured. Hence, if the learning tasks are easy, the test items should be, too. We do not want to modify item difficulty or eliminate easy items from a criterion-referenced test in order to obtain a range of scores (Wiersma & Jurs, 1990). If the instruction has been effective we would expect all or nearly all of the students to obtain high scores. Item difficulty is important but more in the sense of matching the item difficulty to the learning task described in the intended outcome.

In terms of utilizing norm-referenced tests, because we are trying to rank students in order of achievement, deliberate attempts are made to obtain a wide spread of scores. This can be accomplished by eliminating easy questions that everyone is likely to get right or hard items that most people will get wrong, and concentrating on items that maximize the differences in the students performances. To achieve the maximum differentiation in terms of student achievement we want the average score (the mean) to be the midpoint of the possible scores. And we want scores ranging from near zero to near perfect. For example, on a supply-format test (e.g., short answer or fill-in-the-blank) with 100 items, we would want a mean of 50 and a range of scores from 5 to

95.

Advanced Stages in the Construction Process

Although some researchers (e.g., Thorndike et al., 1991; Gullickson & Ellwein, 1985) may have painted a grim picture with regard to teacher-written tests, the fact remains that locally developed measures can be an extremely effective component of teaching and have many advantages. Teacher-written tests can be tailored to the specific needs of the class, they can be administered frequently and, something that sometimes seems to be overlooked, they can assist teachers in identifying individual learner's needs (Worthen, Borg, & White, 1993).

For heuristic purposes we have been discussing tests using a paper-and-pencil format which, of course, may not be best for some types of tests. This leads us to several additional things that must be considered before actually constructing an initial pool of items. First, we must consider the characteristics of the group being tested and how we will test them. For example, young children or children with a learning disability may need to take an oral test in order to obtain reliable scores. There are some practical considerations as well, such as the time needed to administer and score the test and the cost of developing, producing, and administering the test. Finally, we must consider the qualifications of the individuals who will administer, score and interpret the test. This is an especially important consideration because, as we noted earlier, subtle individual differences can effect score validity and the amount of error in

the scores we obtain.

Having followed the afore mentioned procedures, we have now reached the point of constructing the initial pool of items for the test. Again we must consider what will be the best match for the intended purpose. We will look at the advantages and disadvantages of true/false, multiple choice, matching, short-answer, and essay tests.

Item Types

True/False Tests

True/false tests have always been popular with local test developers because they are easy to construct and easy to score. On the downside, these tests may encourage rote learning and may be only testing students at the first level of the cognitive domain. True/false can also expose students to erroneous ideas. Unless the test is fully reviewed with the students after administration, students may "learn" a false statement from the test. Another disadvantage is that true/false tests are susceptible to inflated scores due to guessing.

Sax (1989) offered some guidelines for the construction of the true/false test. First, avoid irrelevant difficulty (don't say ambulate when you can say walk). Second, avoid most negative statements and all double negatives. Third, avoid giving clues to the answer. For instance, using words like "all", "never", and "none" should be avoided because they are associated with false statements. Test writers should also avoid using words like "usually" and "generally" because they are associated with true

items (Gay, 1980).

A good true/false item should relate to a single idea and it should be definitely true or definitely false. For example, using an item that states, "The sun rose yesterday," could be trouble because technically the sun does not rise, the earth moves (Sax, 1989). It is also important to use a random order (no patterns to the answers) and have an equal number of true and false items. False statements are harder to write so there usually are fewer false items on a novice test writer's test.

Multiple-Choice Items

Kubiszyn and Borich (1993) noted that contrary to popular belief, good multiple-choice questions can be "the most time-consuming kind of objective test items to write" (p. 90). Multiple-choice items consist of two parts: a stem and a number of alternatives (Sax, 1989). The stem is a statement or a question that can be answered or completed by chosing one of the alternatives. All of the incorrect or less correct alternatives for the stem are called distractors. The test taker is asked to chose the "best" or "most correct" alternative to complete the stem.

Because there are several types of multiple-choice tests, they are versatile and have numerous advantages. Measurement can be done at all levels of the Taxonomy and, because minimal writing is involved, a good deal of material can be sampled on one test. Multiple-choice items are also easy to score objectively and are particularly amenable to item analyses (Sax, 1989). The fact that

this format is so amenable to item analyses is vitally important to us as test writers because item analyses will allow us to detect areas of student weakness, evidence of item ambiguity, and evaluate item difficulty and the extent to which each item can measure individual differences (Sax, 1989).

Childs (1989) recommended the following guidelines for multiple-choice question construction:

1. State clearly in the instructions whether you require the correct answer or the best answer to each item.

2. Instead of repeating words in each alternative, include these words in the main body of the questions. This will make the question easier to read and the options easier to compare. The grammar and structure of the main part of the question must not contain clues to the correct response however.

3. Make incorrect alternatives attractive to students who have not achieved the targeted learning objectives.

4. Vary randomly the placement of correct responses.

5. Make all choices exactly parallel. Novice test writers tend to make the correct answer longer and more carefully worded and, by doing so, may provide a clue to the correct answer.

6. Never offer "all of the above" or "none of the above" in a best-response multiple-choice question.

> Whether "none of the above" is chosen as a better response than one of the other options may depend on what evidence the student considers rather than how well he or she understands the material.
>
> 7. Control the difficulty of a question by making the alternatives more or less similar or by making the main part of the question more or less specific. If the alternatives are more similar, the student will have to make finer distinctions among them. If the main part is more specific, the student will be required to draw on more detailed knowledge. (p. 2)

There is another important concept with regard to multiple-choice items that we must also consider, the concept of response set. This is not so much a problem with achievement tests, but when constructing self-report inventories, response set can be become real problem. Aiken (1976) defines "response" set as a tendency for test takers to respond in a fixed or stereo-typed way when items consist of two or more possible response choices.

There are two types of response set that may occur in self-report inventories, acquiescence and social desirability. Acquiescence deals with a test taker's tendency to agree with a statement when they have no informed basis for agreeing or disagreeing. An example of this type of response set would a supervisor who fills out an evaluation of a counseling student's counseling skills and responds positively to an item like "HANDLES CRISIS SITUATIONS WELL" when the student had not had any crisis

situations to handle.

Social desirability deals with the test taker's tendency to rate items that are socially desirable with more frequency than items deemed socially undesirable. For instance, if the test item gave the choice of answering "yes" or "no" to "I DRESS LIKE A SLOB," there would be a tendency for the test taker to answer "no." This type of response set can be minimized with the use of a forced choice format. Now the item might read "I PREFER TO DRESS a. formal b. casual c. in whatever I can find."

Matching Items

In many respects, the matching test is really just a type of multiple-choice test in which the test taker associates an item in one column with a choice in a second column. The test taker may associate names of individuals with their accomplishments, events with dates, or countries with their capitals (Sax, 1989). Although the matching format is easy to construct, novice test writers may find it difficult to design items that measure students abilities beyond the first level on the Taxonomy (Knowledge). However, this format is useful for measuring associations and reduces the effects of guessing.

Sax (1989) made the following suggestions for constructing items for the matching test. First, it is important to use homogenous options and items to reduce the possibility of guessing. For example, if the items in a matching set include both people and places, the test taker can easily eliminate certain options for each of those items by matching "people items" with

"people options" and "place items" with "place options."

Related to this issue is a second issue that has to do with the use of specific determiners. Items that contain specific determiners should be avoided because they provide clues for the correct option. Sax (1989) used the example of a matching item that asked for the founder of Pennsylvania. Because the item contains a clue to the correct option (William Penn), it would be easy for any student to guess the correct answer. In this particular case, Sax (1989) suggested adding the choice of "none of the above" to help remedy the problem. Other suggestions for constructing matching tests include arranging options alphabetically or numerically with the shorter responses in the second column and using more options than item stems.

Completion and Short-Answer Items

Short-answer items require students to provide their own answers rather than selecting them from given lists. This format eliminates some of the possibilities for guessing but short-answer items are subject to alternative wordings or long responses as examinees attempt to answer the item correctly. To avoid these problems, Kubiszyn and Borich (1993) made the following suggestions. Omit only key words from completion items and make sure the content of the item is not distorted by the omission. Avoid using direct quotes from textbooks which might promote rote memorization. Also, test writers can lessen the likelihood of alternate or wordy responses by requiring a brief and definitive answer that occurs near or at the end of the item statement.

Essay Items

Essay items have several advantages in educational settings. These items permit us test students at higher levels on the Taxonomy of cognitive skills. They are also easy to construct and are appropriate for small groups of students. But there are disadvantages to the essay format as well. Scoring these items can become very subjective and may also be very time consuming (Sax, 1989). Worthen et al. (1993) stated that some of the broad interpretation and subjectivity can be avoided in the scoring of essay items by constructing questions that are direct, brief, and have a narrow focus. Further, it was suggested that specific instruction regarding time limits and amount of information expected should be communicated to the examinees. Sax (1989) suggested that, if possible an instructor should reread the items or have a peer read them before assigning a final grade.

Revision of Items

Once we have assembled the initial pool of test items, we can begin the process of revision of the items. This can be done by using a review panel of colleagues who are knowledgeable about the subject matter and about test construction. The panel would assess the items for accuracy (appropriateness in terms of age, grade level, subject matter), technical flaws, grammar, offensiveness, and readability. This is the point at which a concept discussed earlier, that of having more questions than will actually be used on a test, comes into play. After a revision, some questions are probably going to be eliminated. Aiken (1976) suggested writing

about 20% more items than will actually be used.

After the initial revision is complete, a pretest should be performed followed by further test revision. Mathieu (1997) suggested conducting the pretest procedure in the following manner. Administer the test to a small sample of examinees (usually 15-30 people). During the administration of the test, assess the reactions of the examinees during the test. Some specific behaviors to watch for among the examinees would be long pauses between responses, scribbling, or changing of answers. Next, invite comments from the examinees once they have finished the test and ask them if they have suggestions for improvement.

Item analysis can also be conducted at this point in the process. Specific things to look at in the item analysis include item difficulty (in terms of the percentage of examinees who got the correct answer) and item discrimination power (the extent to which the item is answered correctly more often by those who obtained higher test scores than by those who obtained lower test scores (Wiersma & Jurs, 1990). Upon completion of the item analysis further item revisions can be made.

Conclusions

As can be seen from this brief overview of basic test writing precepts and principles, writing good tests can be difficult and time consuming. However, the stakes in the test game are high and the potential rewards in both education and research are great. In education, we must always remember that the purpose of testing is not only to assess what students have learned, but also to help us

teach more effectively and, ultimately, to help students to master more of our course objectives (Childs, 1989). In research, the integrity of our assessments and research endeavors turns upon the quality of our tests.

References

Aiken, L. R., Jr. (1976). Psychological testing and assessment (2nd ed.). Boston: Allyn and Bacon, Inc.

Bloom, B. S., Englehart, M. D., Furst, G. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: Handbook I, The cognitive domain. New York: David McKay Co., Inc.

Brown, F. G. (1983). Principles of educational and psychological testing (3rd ed.). New York: Holt, Rinehart and Winston.

Childs, R. A. (1989). Constructing classroom achievement tests (Digest EDO-TM-89-8). Washington, DC: American Institutes for Research. (ERIC Document Reproduction Service No. ED 315 426)

Crocker,L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Gronlund, N. E. (1971). Measurement and evaluation in teaching (2nd ed.). New York: The Macmillan Company.

Gulliksen, G. (1986). Perspective on educational measurement. Applied Psychological Measurement, 10, 109-113.

Gay, L. R. (1980). Educational evaluation and measurement: Competencies for analysis and application. Columbus, OH: Charles E. Merrill Publishing Co.

Kubiszyn, T., & Borich, G. (1993). Educational testing and measurement (4th ed.). New York: Harper Collins College Publishers.

Mathieu, C. K. (1997). Basic precepts in test construction: Recommendations from various measurement textbooks.(ERIC Document Reproduction Service No. ED 407 424)

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: Macmillan.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Wiersma, W. & Jurs, S. G. (1990). Educational Measurement and testing (2nd ed.). Boston: Allyn and Bacon.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). Measurement and evaluation in the schools. New York: Longman.